웹 사용 마이닝을 위한 SASOM+DT를 이용한 웹 데이터의 분류

유시호[0], 김경중, 조성배
연세대학교 컴퓨터과학과
{bonanza[0], uribyul}@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Classification of Web Data Using SASOM+DT for Web Usage Mining

Si-Ho Yoo[0], Kyung-Joong Kim, Sung-Bae Cho
Computer Science Department, Yonsei University

요약

웹 마이닝은 크게 구조 마이닝, 컨텐츠 마이닝, 사용 마이닝으로 분류될 수 있다. 이 중에서도 사용 마이닝은 사용자의 로그 데이터를 바탕으로 사용자가 탐색한 웹 페이지의 순서를 추출하거나 연관관계를 분석하는 작업이다. 특히 웹에 기반을 둔 애플리케이션의 요구를 충족시키기위해서 사용 마이닝은 웹 마이닝에 있어서 중요한 부분으로 부각되고 있다. 본 논문에서는 사용자들의 웹 페이지의 방문 패턴을 분석하여, 미래행동을 예측하는 것을 문제로 삼고, 사용자들의 이용패턴을 SASOM(Structure-Adaptive SOM)분류기들의 DT(Decision Tree)앙상블을 이용하여 분류하는 방법을 제안해보았다. MS웹 데이터를 가지고 SASOM 분류기의 집합을 DT를 이용하여 결합한 결과, 분류기 하나만 사용한 경우보다 더 좋은 결과를 얻어, 3.5% 이하의 낮은 오류율을 보였다.

1. 서 론

최근 들어 웹상에서 행해지는 각종 비즈니스 거래는 엄청난 속도로 증가하고 있으며, 전자 상거래란 신단어가 나올 정도로 현존하는 많은 기업들이 웹을 이용한 거래에 상당부분 의존하고 있다. 특히, 웹 거래의 특성상 한쪽 끝에 존재하는 고객들이 컴퓨터를 사용하는 사용자들이기 때문에 이러한 개인의 정보는 전자상거래에 있어 매우 중요하다. 사용자의 웹 탐색 행동을 추측하는 능력은 상인과 고객간의 거리를 굉장히 가깝게 만들기때문에, 각 고객의 취향에 알맞은 상품정보를 다양하게 제공하는 방법이 중요해졌다.

웹 데이터를 가지고 사용자들의 사용패턴을 분석하는 웹 사용 마이닝은 다양한 분야에서 응용되고 있다. 본 논문에서는 웹 사용 마이닝을 위하여, 사용자들의 사용패턴으로부터, 미래 행동을 예측하는 것을 문제로 삼았다. 각 개인마다 웹 사이트를 방문하는 순서, 경향이 다르므로 이를 분석하여, 미리 사용자들의 방문여부를 예측하는 것 또한 전자거래에 매우 중요한 부분이라고 할 수 있다. 이를 예측하기 위해서 분류기들의 앙상블 방법을 제안하였다. 앙상블이란 여러 개의 분류기들의 집합으로, 각각의 분류기를 서로 다른 학습데이터로 훈련을 시키고, 학습된 분류기들의 결과들을 결합하는 방법을 말한다. 이 방법은 한 명의 전문가보다는 여러 전문가의 견해를 듣고 판단을 하는 것이 도움이 더 되듯, 서로 다른 여러 개 의 분류기를 사용하여 결합하는 방법이 분류기 자체의 성능을 향상시킬 수 있다는 전제하에서 출발한다.

2. 웹 사용 마이닝

웹 마이닝에서의 데이터는 서버, 클라이언트, 프록시서버 또는 데이터베이스에서 추출될 수 있다. 이런 데이터는 그들의 소스 위치뿐만 아니라, 데이터의 유용성, 함축성 등에서 다르다. 이러한 데이터를 가지고 마이닝 작업을 하는데, 웹 사용 마이닝을 하기 위해서는 세 가지 단계적인 작업이 필요하다. 전처리 작업, 패턴 발견 작업, 패턴 분석 작업이 단계적으로 실행된다[1].

2.1 전처리

전처리는 적절한 데이터의 용도, 내용 그리고 구조정보를 패턴 발견 작업에 필요한 형태로 변형시키는 작업을 뜻한다. 수집된 데이터를 사용자 또는 관리자에게 유용한 정보를 제공하기 위한 형태로 변환하기 위해 연관, 클러스터링, 순차 패턴, 경로 분석 등의 작업을 거친다.

2.2 패턴 발견

패턴 발견은 통계학, 데이터마이닝, 기계 학습, 패턴 인식 등의 분야에서 활용했던 알고리즘들을 이용한다. 이 중에서도 특히 클러스터링, 분류, 통계적 분석이 가장 많이 사용되고 있다. 이러한 방법들은 웹 마이닝에 필요한 사전 지식과 데이터의 용도 등을 고려하여 선택된다.

2.3 패턴 분석

웹 사용 마이닝의 마지막 단계로서 패턴 발견 작업에서 발견한 패턴들 중에서 유용하지 않은 패턴이나, 규칙 등을 제거하는 작업을 한다.

3. SASOM+DT

웹 사용 마이닝에서 두 번째 단계인 패턴 발견 작업은 다양한 방법을 통해 진행되는데, 본 논문에서는 앙상블을 사용하는 방법을 제안하였다. 분류기의 앙상블을 사용자들의 패턴을 입력하여 훈련시키고, 이것을 이용하여 사용자들의 패턴을 예측해보았다. 현재까지 개발되었고, 사용되고 있는 분류기는 매우 많지만, 실험에는 코호넨이 개발한 SOM을 변형시킨 SASOM[4]을 사용하였다.

3.1 자기구성지도

SOM은 자기구성 지도로 불리며, 튜보 코호넨이 제안한 신경망으로 어떻게 자기 스스로 구조를 생성해 낼 수 있는 지에 대한 답을 비교적 간단하게 제시해 주고 있다. 자기 구성이라는 말은 신경망이 주어진 입력에 대해 올바른 출력 값이 제공되지

않고도 학습됨을 의미한다[2].

SOM은 결과를 미리 알지 않은 상태에서 자기 스스로 학습을 할 수 있으므로 다른 신경망 들에 비해서 생물학적인 특성에 잘 부합한다고 볼 수 있다. 그러나 이런 규칙을 가지고 학습되는 SOM은 하나의 출력노드에 다수의 클래스가 분포하는 경우가 생기는데 이런 경우 올바르게 분류하기가 힘들어진다. 그렇기 때문에 이러한 노드들을 다시 더 작게 분화시키는 방법이 필요한데 그것이 SASOM이다.

3.2 구조적응 자기구성지도

구조적응 자기구성지도(Structure Adaptive SOM)라 불리는 SASOM은 자기구성 지도에서 노드를 더 작게 분화시키고, 학습시킨후에 불필요한 노드들을 삭제하는 기능이 추가된 자기구성지도이다[3,4].

기본 알고리즘은 다음과 같다.

[1] 지도를 4×4 크기로 초기화한다.
[2] SOM 알고리즘으로 학습시킨다.
[3] 지도의 노드들 중 여러 클래스의 데이터가 섞인 노드를 찾는다.
[4] 단계3에서 발견된 노드들을 2×2크기의 노드로 분화시킨다.
[5] 분화된 노드들을 LVQ알고리즘으로 학습시킨다.
[6] 학습에 참여하지 않는 노드를 삭제한다.
[7] 종료 조건이 만족될 때까지 위의 단계를 차례대로 반복한다.

그림 1. SASOM 알고리즘

이 알고리즘은 지도 초기화 및 초기 학습단계, 노드 분화단계, 분화된 노드 학습 단계의 세 부분으로 나뉘어진다.

(1) 지도 초기화 및 초기학습 단계

지도의 크기를 임의로 설정하여 초기화하고 코호넨 알고리즘에 의해서 학습시킨다. 지도는 4×4의 크기로부터 시작하여 학습된다.

(2) 노드 분화 단계

이 단계는 초기화된 지도를 토대로 분화되어야 할 노드를 찾아내는 역할을 한다. 노드가 하나의 클래스에 대해서 반응하는 것이 아니라 다수의 클래스에 대해서 반응하게 되면 잘못된 결과를 산출하기 때문에, 이러한 노드들을 찾아서 분화시켜야 한다. 분화된 하위 노드들의 가중치는 분화되기 전 부모 노드의 가중치 값을 기반으로 이웃한 노드들의 가중치 값들을 고려하여 다음의 식과 같이 산출된다.

$$C = \frac{(P \times 2) + \Sigma Nc}{S} \quad (1)$$

여기서, C는 자식 노드이 가중치, P는 부모 노드의 가중치, Nc는 자식 노드의 이웃노드의 가중치, S는(Nc의 개수+2)를 나타낸다. 즉 C는 이웃노드와 부모노드의 평균값으로 결정된다.

(3) 분화된 노드 학습 단계

분화가 일어난 노드들은 학습을 시켜야 한다. 이 단계에서의 학습은 순수한 코호넨 알고리즘에 LVQ알고리즘을 결합한 형태의 교사 학습 형태를 가진다. 학습식은 다음과 같다.

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\} \quad (2)$$

여기서 $n_{ci}(t)$는 최상매칭 클래스에 대해서는 가중치를 증가시키고($n_{ci}(t)$=1) 그 이외에는 가중치를 주지 않고 학습을 시키지 않는다($n_{ci}(t)$=0). 이와같이 SASOM은 SOM으로는 분류할 수 없었던 노드들을 분화시켜 보다 정교한 분류기능을 수행할 수 있게 된다.

3.3 DT에 의한 결합

본 논문에서 사용한 앙상블 모델은 SASOM의 조합으로 기본 알고리즘을 살펴보면 다음과 같다.

```
while(more data points are available)
  while(all classifiers in ensemble are not trained)
    read d points, creating training set D
    train classifier C with D
  end
  train extra classifier Ce with new D
  evaluate all classifiers with test data
  if(quality(Cj)<quality(Ce) for some j)
    replace Cj with Ce
end
```

그림 2. 앙상블 알고리즘

각 분류기는 훈련데이터로부터 일정한 크기(d)로 자른 순차적인 데이터(D)에 의해 훈련된다. 데이터의 크기는 임의로 고정시키며, 앙상블내의 분류기의 개수도 임의로 고정시킨다. 만약, 데이터의 크기(d)를 100개로 하였다면 먼저 훈련데이터로부터 100개의 데이터를 입력받아서 하나의 SASOM(C)을 만든다. 그리고 첫 번째 100개의 데이터의 끝부터 시작하여 또 다른 100개의 데이터를 훈련데이터로부터 입력받아서 두 번째 SASOM을 만든다. 이렇게 앙상블내의 모든 SASOM를 생성한다. 이때 생성되는 모든 SASOM들은 서로 다른 데이터로 훈련되었기 때문에 전부 다르다. 훈련이 끝나고나면 훈련데이터로부터 다시 100개의 데이터를 입력받아서 또 하나의 SASOM(C_e)을 만든다. 그리고 테스트 데이터로부터 역시 같은 크기(d)의 데이터 100개를 입력받아서 지금까지 훈련시킨 모든 SASOM을 테스트한다. 테스트한 결과들을 비교하여 만약에 추가적으로 만든 SASOM(C_e)이 앙상블내의 특정 SASOM 보다 성능이 좋다면, 서로 교체해준다. 앙상블내의 모든 SASOM보다 추가적으로 만든 SASOM(C_e)의 성능이 나쁘다면, 교체하지 않고, 훈련데이터로부터 다시 100개의 데이터를 입력받아서 또 다른 SASOM(C_e)을 만들고 위의 과정을 반복한다. 그림 2의 알고리즘은 이와 같은 방법으로 진행된다[5].

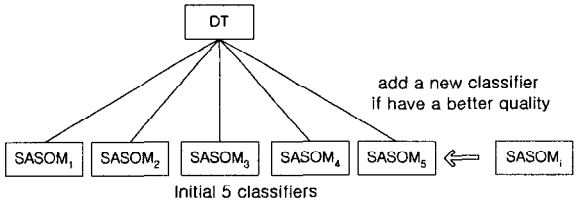

그림 3. 5개의 SASOM을 사용한 앙상블

그림 3은 5개의 SASOM을 가진 앙상블을 나타낸 것이다. 이와같이 알고리즘이 끝날 때까지 새로운 분류기를 계속 만들어 추가할지를 결정하고 마지막에 DT(Decision Tree)로 그 결과들을 결합한다[6].

4. 실험 및 결과

실험 데이터로는 UCI KDD 데이터베이스의 Anonymous Web Browsing Data를 사용하였다. 이 데이터는 총 32,710명의 방문자가 1998년 2월중 한 주간 Microsoft web site를 방

문한 기록을 가지고 있다. (http://kidd.ics.uci.edu/) 이 데이터는 방문자가 총 294개의 사이트를 방문한 여부에 대한 기록을 가지고 있다.

294개의 사이트 중에서 방문자가 'free download' 페이지를 방문하였는지를 예측하는 것을 문제로 삼고 실험을 하였다. 'free download' 를 방문하였으면 클래스를 1, 방문하지 않은 경우는 클래스를 0으로 하여 학습데이터 32,710개를 가지고 훈련을 시키고, 테스트데이터 5,000개를 가지고 제대로 예측하였는지를 평가하였다.

앙상블내의 분류기의 개수는 5, 10, 15개로 증가시켜서 실험하였고, 입력받는 데이터의 크기도 100, 200, 500으로 증가시키면서 평가하였다.

표 1. 앙상블의 오류율

	100 data	200 data	500 data
5개의SASOM	6.5%	5.0%	3.5%
10개의SASOM	4.0%	3.5%	2.5%
15개의SASOM	2.6%	1.8%	0.8%
20개의SASOM	3.0%	1.6%	1.0%

표 2. 1개의 SASOM으로 분류한 오류율

	100 data	200 data	500 data
1개의 SASOM	4.0%	5.5%	3.8%

표 1은 앙상블에 대한 결과값으로 전체 데이터에 대해 잘못 분류한 데이터의 비율인 오류율을 나타내었다. 데이터의 크기를 증가시킬수록, 앙상블내의 분류기를 증가시킬수록 오류율이 작아지는 것을 볼 수 있다.

표 2는 학습데이터로 훈련시킨 하나의 SASOM을 가지고 분류한 오류율을 나타낸다. 표1과 비교 해보면, 데이터가 작거나, 분류기의 개수가 적을때는 앙상블과 하나의 SASOM은 별 차이가 없지만, 데이터가 많아질수록, 분류기의 개수가 많아질수록 하나의 SASOM보다는 앙상블의 성능이 더 좋아짐을 알 수 있다.

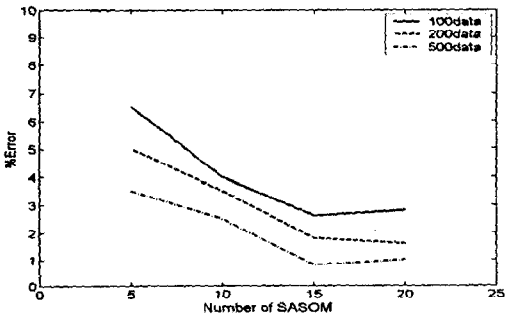

그림 4. 앙상블의 에러율

그림 4를 보면 쉽게 분류기의 개수와 데이터의 크기가 앙상블 전체의 성능과 어떠한 관계가 있는가를 알 수 있다. 데이터의 크기는 500개, 앙상블내의 분류기는 15개일때가 앙상블의 오류율이 가장 낮다. 하지만 분류기의 개수가 20개 이상으로 넘어가면, 앙상블의 오류율은 별 차이가 없기 때문에 분류기의 개수와 앙상블의 성능과는 반드시 비례관계가 아님을 알 수 있다.

그림 5. 분류한 지도

그림 5에서 왼쪽에 있는 지도가 5개의 SASOM으로 500개의 데이터를 테스트한 결과이고, 오른쪽에 있는 지도가 15개의 SASOM으로 500개의 데이터를 테스트한 결과이다. 두 개의 지도 모두 노드분화를 끝낸 최종지도이다. 초기에 4×4크기의 지도로 시작하여 최종적으로 그림 5와 같은 형태의 지도로 분화된 것이다. 15개의 분류기로 형성된 지도가 훨씬 더 세밀한 원으로 잘 분화된 것을 쉽게 알 수 있다. 5개의 분류기를 사용한 지도를 보면 처음 크기 그대로의 노드들도 상당수 있는 것으로 보아, 분화가 잘 안된 데이터가 많은 것을 알 수 있다.

본 논문에서 사용한 앙상블 방법은 하나의 훈련데이터를 가지고 임의의 크기로 자른 서로 다른 데이터로 분류기들을 훈련시켰다. 이는 기존의 앙상블 방법이 서로 다른 특징으로 추출된 훈련 데이터를 사용한 것과 비교하여 다르다. 방법 자체가 여러 가지 특징 추출을 할 필요가 없다는 점에서 매우 간단하며, 그 결과 또한 단순한 분류기보다 성능이 좋다는 것을 알 수 있다.

참고문헌

[1] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Ting, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 1-2, pp.12-23, 2000.

[2] T. Kohonen, "The self-organizing map," *Proceedings of IEEE*, vol. 78, no. 9, pp.1464-1480, September 1990.

[3] H. D. Kim, and S. B. Cho, "Application of self-organizing maps to classification and browsing of FAQ emails," *PRICAI Workshop*, pp. 44-55, 2000.

[4] S. B. Cho, "Ensemble of structure adaptive self-organizing maps for high performance classification," *Inforrmation Science*, vol. 123, no. 1-2, pp. 103-114, 2000.

[5] W. N. Street, and Y. S. Kim, "Streaming ensemble algorithm(SEA) for large-scale classification," *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.377-382, San Francisco, California, 2001.

[6] S. Gutta, J. R. J. Huang, P. Jonathon, and H. Wechsler, "Mixture of experts for classification of gener, ethnic origin, and pose of human faces," *IEEE Transactions on, Neural Networks*, vol. 11, no. 4, pp. 948-960, July 2000.